SMU Data Science Review

Volume 5 | Number 3 Article 10

2021

Aspect-based Sentiment Analysis of Movie Reviews

Samuel Onalaja
Southern Methodist University, sonalaja@mail.smu.edu

Eric Romero
Southern Methodist University, edromero@mail.smu.edu

Bosang Yun
Southern Methodist University, byun@smu.edu

Aspect-based Sentiment Analysis of Movie Reviews

Samuel Onalaja, Eric Romero, Bo Yun, Faizan Javed

Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA

sonalaja@mail.smu.edu, edromero@smu.edu,
byun@smu.edu, fjaved@mail.smu.edu

Abstract – This study investigates a comparison of classification models used to determine aspect based separated text sentiment and predict binary sentiments of movie reviews with genre and aspect specific driving factors. To gain a broader classification analysis, five machine and deep learning algorithms were compared: Logistic Regression(LR), Naive Bayes (NB), Support Vector Machine (SVM), and Recurrent Neural Network Long-Short-Term Memory (RNN LSTM). The various movie aspects that are utilized to separate thesentences are determined through aggregating aspect words from lexicon-base, supervised and unsupervised learning. The driving factors are randomly assigned to various movie aspects and their impact tied to each aspect and genre leading to the sentiment classification has been fully investigated based on the accuracy of each model. The study shows that assigning higher driving factors to certain aspects and genre result in the higher accuracy of the sentiment prediction models that utilized in this research.

1 Introduction

The emergence of several movie streaming platforms such as Amazon Movies, Netflix, and Hulu have allowed users to share their opinions in several formats efficiently. These include reviews in formats such as text, videos, and rating systems to denote a range of positive to negative sentiments.

To understand customers' preferences and improve customer experiences, businesses use the method of opinion mining to identify the sentiments of textual information. These methods can be complex because of misspelled words, abbreviations, emoticons, and slang. Therefore, it is important to identify the appropriate opinion mining method to produce accurate analysis of users' sentiments. The establishment seeks the help of experts, who apply sentiment analysis to gain insights into the collected data. Sentiment analysis is the computational study of people's opinion of an entity and is one of the most active areas of research.

Sentiment analysis is the process of extracting subjective information from text and it is becoming immensely popular because it is highly efficient, and thousands of reviews can be analyzed in a short time in several ways. It uses machine learning techniques and natural language processing (NLP) to analyze and make statistical inferences from textual information in several ways of applications which include the following:

- Recommendation System
- Product and Service Quality Improvement
- Organizational Decision Making
- Customer Decision Making
- Organizational Marketing Research

The dataset used in this research was obtained by scrapping IMDB for review analysis. Using this data, several models will be built to obtain sentiment analysis on a variety of aspects within several film genres. This paper aims to build classification models that show whether viewers' sentiments are positive or negative. Since traditional sentiment analysis focuses only on classifying sentiments without specifying the parts, Aspect-Based Sentiment Analysis (ABSA) will be utilized in the study. This will help analyze common words, slang, emoticons, and typographical errors related to different movies.

Aspect Extraction (AE) is the process of extracting all the terms related to the aspect of a product/service, such as slang, abbreviations, emoticons, and typographical errors. Performing this task requires a sequence labeling in which each input word is labeled as either B, I, or O, where B means Beginning, I means inside, and O means outside. This is necessary to show the position of the term as aspects can sometimes contain two or more words.

Aspect Classification (ASC) is the process of classifying sentiments according to their categories: Positive, Negative. The task involves scraping review data from the IMDB website, scouring the dataset for missing data or outliers, and making sure the dataset is not biased so we can get a near accurate result and then build models on the cleaned-up dataset. Support Vector Machine, Logistic Regression, Naïve Bayes, and Long Short-term neural networks will be implemented in conjunction with aspect components AE and ASC, and genre driving factors.

Most of the past research and papers were focused on binary classification but this paper will cover more in-depth understanding of sentiment analysis including various aspects of movie reviews such as casts, music, location, technology and quality. This will help in instances where reviewers are interested in quality or cast of a movie or in a case where a reviewer is only interested in watching a movie by a particular producer and wants to read different reviews of their produced movies.

Since previous papers have discussed different methods for sentiment analysis on movie review datasets, this study will supplement existing research of movie review by focusing on Aspect-Based Sentiment Analysis (ABSA) with aspect and genre specific driving factors to further develop granular level of sentiment understanding and prediction. This research will help contribute to existing research on sentiment analysis and Aspect-Based Sentiment analysis (ABSA) for movie reviews and to help make informed business decisions regarding movie qualities and customer satisfaction.

2 Literature Review

As people spend more time watching movies through streaming services, the need also increases for efficient assessment of which movies to recommend. One proven method is sentiment analysis of movie reviews, which have become more available to the public through developments in online media. In the following article "A difference of multimedia consumer's rating and review through sentiment analysis" published by Lee, Jiang, Kong, and Liu in 2020, authors address a strong need for review-based sentiment analysis in the consumer world. "This is so because services are difficult to predict until they are experienced" [23].

2.1 History of Film Criticism

The first film critiques came soon after the dawn of film media in the early 1900s. As films became more popular, newspapers began hiring professional critics to write more serious analysis of the films to add more than just entertainment value [6]. New styles of film analysis developed over time and eventually became a standard feature for prominent magazines.

In more modern times, film critique was additionally made popular through television media. Established critics Roger Ebert and Gene Siskel were notable for developing the show “Siskel & Ebert At the Movies” in the 1980s that would not only review films, but also conduct interviews with film actors. The main task for most review media is to explain a film's premise in addition to its artistic or entertainment merits. Film summaries are often expressed through a rating system such as numeric scales, grades, image representations, or “thumbs” in the case of Siskel and Ebert.

2.2 Development of Online Film Criticism

Online blogs were one of the first internet media to be used for film criticism, allowing any person to write their opinion of a film for others to read. However, audience size was limited by the popularity of amateur writers and the sites they used. Using more modern platforms such as YouTube function in a similar fashion but provide access to a wider audience and interests with the use of videos, cut-scenes, animations, and actors to express film critiques.

Specialized websites were also developed to provide a direct source for film critiques and reviews. Specific types of criticism have developed within online media that focus on particular aspects such as scientific realism, plot holes, and theories on possible sequels. Other sites may be specifically tailored to offer analysis on aspects such as content advisories, for parents concerned with their children watching the film.

Several sites have dedicated their use to providing a source for the general public to express their views on films. These typically incorporate a written commentary from the user that can vary greatly in length depending on depth and breadth. Additionally, a scaled rating system is commonly included that is then used to calculate an average rating and rank to compare with other movies.

The modern film criticism industry has been shown to exhibit some bias, particularly toward gender. Often it is the case that reviews are more male dominated with fewer representations of women. For example, male reviewers authoring articles in Time magazine or radio critics on NPR, have been shown to represent approximately 70-80 percent of those formats. [8] Changes initially introduced by the internet led to a decrease in women working as film critics in newspapers. This eventually developed into shortages of women as opinion columnists overall. Men were more likely to retain these roles and therefore became the more prevalent voice for reviewers. [20]

2.3 History of Sentiment Analysis

Early work in prediction based on sentiment analysis can be found in a study by Hatzivassiloglou and McKeown in 1997. Using a corpus-based approach they analyzed adjective terms in stock market reports to achieve 90% sentiment precision with their model [15]. In 2004, a study by Pang and Lee utilized machine learning to analyze sentiment to select subjective sentences and use them to categorize text. Using a Naive Bayes model, they achieved an accuracy of 86.4% in determining sentence polarity [30]. One of the earliest studies for online comments was in 2005 by Gruhl et al. In the research they analyzed blog commentary to successfully predict changes in book sales data from Amazon [13].

Modern methods for sentiment analysis are now separated into three categories: knowledge-based, statistical, and a hybridized combination of the two [8]. In the knowledge-based methods a lexicon or corpus is used to identify the presence of emotion words such as happy or sad. These are then used to make a classification of the overall text sentiment [29]. Statistical techniques utilize machine learning algorithms to classify text as a range of positive to negative [37].

2.4 Studies of Machine Learning tools in Sentiment Analysis

Research has continuously focused on methods for extracting meaningful information from reviews, while also being able to categorize the information into positive or negative sentiment. In a 2018 publication, Nguyen, Veluchamy, Diop, and Iqbal completed a comparative study between machine learning and lexicon-based models for classifying text sentiment. Amazon product reviews were used to predict 1-5 star ratings where accuracy, precision and recall were measured and resulted in the machine learning models generally performing better. One issue that had to be dealt with for future consideration was the use of emojis in the text which had to be removed for processing, this would be a consideration in all types of reviews. Their best result was using a logistic regression model which performed at a reported 90% accuracy [27].

Use of logistic regression was also common to another publication where Yelp review features were extracted from generated scores using text classifiers and sentiment analysis. It was found as part of the study that factors leading to the classification of a recommended review were far easier to classify when "written in a few moderately complex sentences expressing substantive detail with an informative range of varied sentiment." This information would be something to consider in our own analysis of movie reviews to obtain higher quality text sentiment. In the logistic model that was created, features were quantified as coefficients in a logistic regression model that was able to predict whether a review would be recommended at an accuracy of 78% [41].

Naive Bayes is also commonly selected for text-based classification. This is also widely used when high dimensionality is an issue within the dataset which is a common occurrence with any language processing and lexicon development. In several studies Naive Bayes outperformed several machine learning models like SVM and logistic regression indicating this should be considered for use as our lexicon is likely to become quite large given the size of each review in consideration [34].

This was similarly true in a similar study using Naive Bayes against KNN, Decision Tree, and Random Forest models on sentiment analysis including a 5-star rating system similar to the dataset used in this research [38][39]. It was also suggested that hybrid methods between models may become more effective than Naive Bayes alone and would be worth investigating to further improve model accuracy [5].

Naive Bayes also showed improved results in accuracy when the original lexicon

was built using a unigram, bigram and trigram patterns. These included negative words and intensive adverbs used as features. An improved Naive Bayes algorithm was introduced to solve imbalances in positive classification and negative classification accuracy which may prove useful for distinguishing between review ratings within this research [4][18].

In a 2018 publication Khaleghi, Cannon, and Srinivas evaluated commonly used hotel review recommender algorithm methods by comparing their accuracy. Collaborative filtering is a method in which predictions are made based on user ratings using the theory that users with similar ratings will like similar things. Matrix factorization uses "single value decomposition," which is typically more accurate except in the case of sparse data sets. In the experiment they found a significantly better accuracy for the matrix factorization but at a cost of 10 times longer processing speed. In addition, they also concluded that the limited size of their sampling data for linking commonalities between users was affecting the outcomes. This procedure can be emulated for the purposes of this research. Data obtained from specific professional reviewers that typically make reviews on all types of movies would be more easily separated to avoid sparsity [19].

2.5 Studies in Aspect-based Models

After discussing the importance of sentiment analysis and various techniques utilized above, Document-level Sentiment Classification seems a reasonable measure of classifying textual information into negative or positive sentiment traditionally using lexicon-based techniques or simple machine learning algorithms. However, simply aggregating the polarity of each word in the review to get the binary sentiment, does not quite provide enough insights and can potentially overlook different mixtures of sentiments that negate each other [31]. As viewers are prone to have different opinions about different aspects of the movie, it is better to examine different aspects of the review. The aspect-based model will be able to predict positive and negative binary sentiment of a written movie review using aspect-based sentiment analysis of the reviewer text combined.

2.6 Data Selection

It is imperative to take a precautionary approach when choosing the right dataset for this study, especially for movie reviews that are generated from online users. Biased algorithmic decision making will treat people of color, race, and nationality unfairly and lead to discrimination and marginalization [1]. To mitigate the bias and promote fairness of the study, the two most popular movie review platforms were compared to choose the dataset that is more balanced and representative of the general audience: IMDB and Rotten Tomatoes

The two platforms use different score metrics. IMDB uses weighted average of its registered users' votes to calculate the ratings [16]. And this methodology prevents extreme values from affecting the result in what is presumably an asymmetrical distribution [17]. Tomatometer from Rotten Tomatoes uses a percentage figure, which is simply a percentage of the critics who rated positive. Tomatometer has a systematic bias that explains why it does not perform well in predicting the box office sales [17].

IMDB assists better with knowledge in general audience than Rotten Tomatoes with its ratings breakdown in user demographics. IMDB shows distribution among age and gender although geotagging is limited to U.S and Non-U.S. None of the demographics data is available in Rotten Tomatoes. Rotten Tomatoes is more gender biased than IMDB: a study on top critics on Rotten tomatoes are 91% male [9], whereas IMDB shows 70% male [36].

General audience seeks entertainment with a good story that fulfills their satisfaction. Professional critics look to dissect the movies with their own critical standards. IMDB reflects general audiences' views while Rotten Tomatoes reflects a small group of selected critics. Overall, IMDB dataset is more representative of the general audience, whose reviews dictate the decision of both ordinary and avid movie goers, thus is concluded to be more suitable than Rotten Tomatoes' for this study.

3 Data

3.1 Web Scraping

An API request was made on IMDB to get 3000 movie titles (600 movies reviews per genre). Using the "BeautifulSoup" package, website URLs that link to each movie review page were scraped. For each movie, a positive and a negative review were extracted using minimum and maximum score. This ensures that reviews are clear with sentiments on either side. For each review, the genre of the movie was imported as well. A total of 3,000 movie reviews of equal number of positive and negative reviews were collected. Movies will be scraped and collected in a manner that the genre will be equally distributed among Horror, Comedy, Action, Romance and Sci-fi. The resulting data frame is then randomly shuffled to be fed into the preprocessing for the models.

3.2 Text pre-processing

It is crucial to keep the dimensionality of the text low to improve the performance of the machine learning classifier [14]. Thus, it is highly recommended to remove the noise as much as possible and to properly preprocess the text in this pre-analysis stage.

Contractions: Dictionary of contractions is provided to be filtered in each review and contractions will be expanded. All the capital letters will be converted to lowercase.

Punctuations and special characters: and html special characters will also be removed.

Tokenization: Each review will be separated into smaller units called tokens using NLTK package tokenization.

Stopwords: will be removed using nltk package's built-in function.

POS tagging: A generic POS tagging is applied to classify words into four categories of adjective, verb, noun, and adverb. More detailed tagging techniques will be developed to further increase accuracy of lemmatization. An alternate POS tagging technique will be an algorithm that resembles a voting engine will be developed by combining several different POS techniques. The POS tagging that the majority picked will be selected.

Lemmatization: The reviews will be lemmatized based on the POS tagging so it is crucial to have accurate tagging classification.

3.3 EDA

3.3.1 General sentiment

Average number of adjective words per rating distribution and most common words for each sentiment have been studied. The following graph shows a normal distribution of sentiment and subjectivity. Lemmatized data set was then plotted by the most frequent adjectives in each sentiment to show the separation in word representations.

3.3.2 General sentiment distribution

The following chart is sentiment distribution measured by the Textblob package. Textblob is a sub-package of NLTK and can measure textual polarity of either -1 and 1 (negative and positive) based on predetermined semantic labels. The sentiment distribution is mostly normal with a slight skewness to the positive sentiment as in Figure 1. It also measures the subjectivity of the text and how much it contains opinions. Most of the text shows a good balance between opinions and facts.

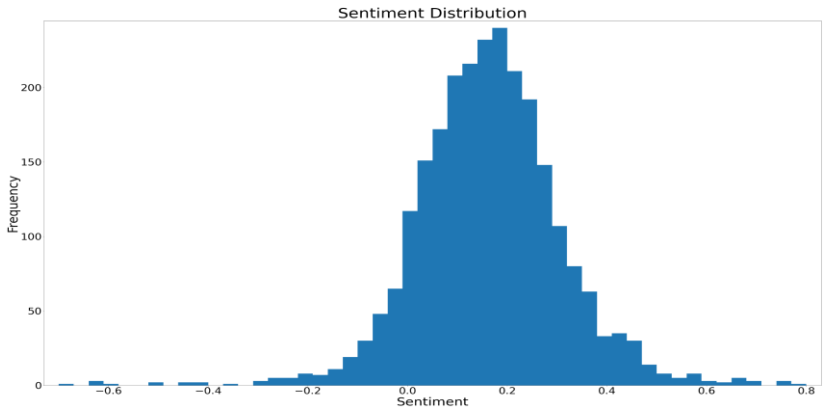

Fig. 1. Subjectivity distribution is displayed. It is a normal curve with slightly right skewness

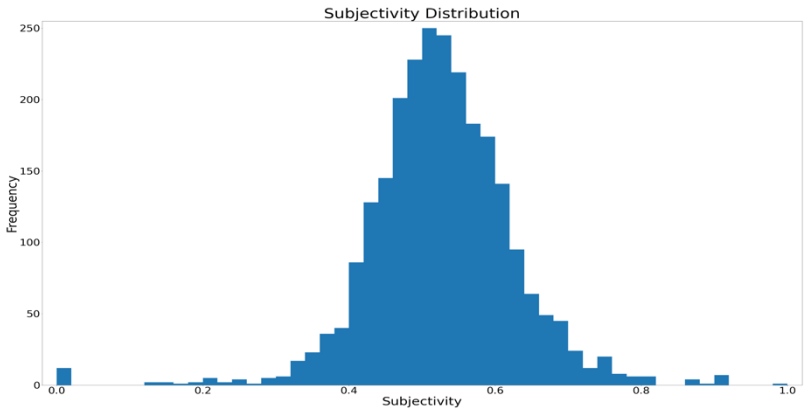

Fig. 2. Subjectivity distribution is displayed. It is a normal curve with slightly right skewness

3.3.3 Word Cloud

As expected from the normal curve pictured above, the words are slightly skewed toward the positive sentiment as shown in the word cloud below. More refined word preprocessing will be implemented and see how this composition might change accordingly.

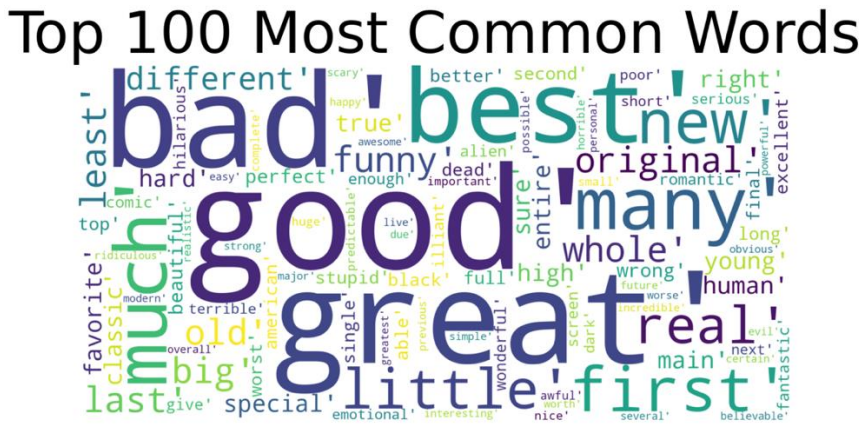

Fig. 3. Word cloud shows a mixture of positive sentiments.

3.3.4 Aspect based sentiment

After dominant aspects are identified in each review, negative and positive occurrences of each dominant aspect are counted and plotted for EDA purposes. Using the timestamp of each review, the count of the most frequent words for each year will be measured and plotted for the trend analysis over time. After the top five most and least frequent words are identified for each year, they will be put in a lexicon list to be assigned special weights (frequent words for less weights, least frequent words for more weights).

3.3.5 Genre-specific EDA

Each genre is examined to see what most frequent words belong to them. Even though comparing two opposite genres "Comedy" and "Horror" were expected to have clear distinction in words, the general words and their frequencies were similarto each other. The similarity in lexicon list between the two genre underscores the necessity of applying aspect-based analysis for granular level of sentiment understanding in different aspects and genre.

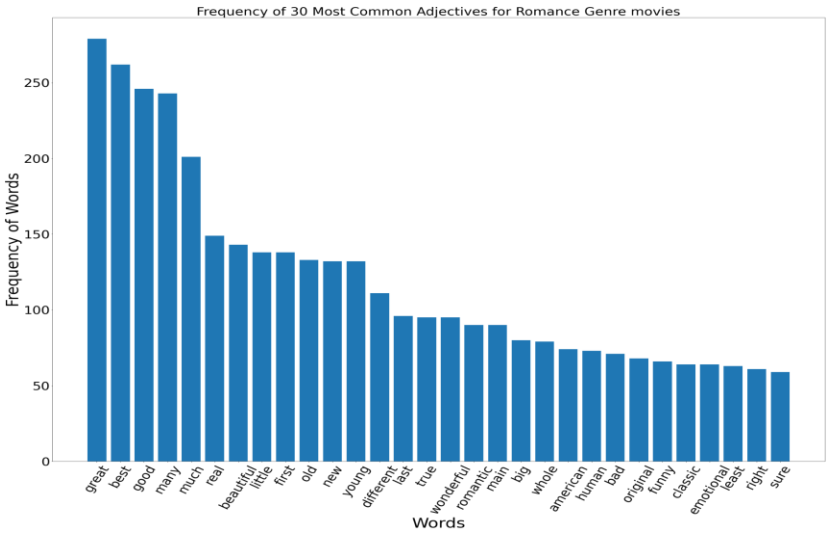

Fig. 4. Top 30 most frequent words shown for Romance Genre

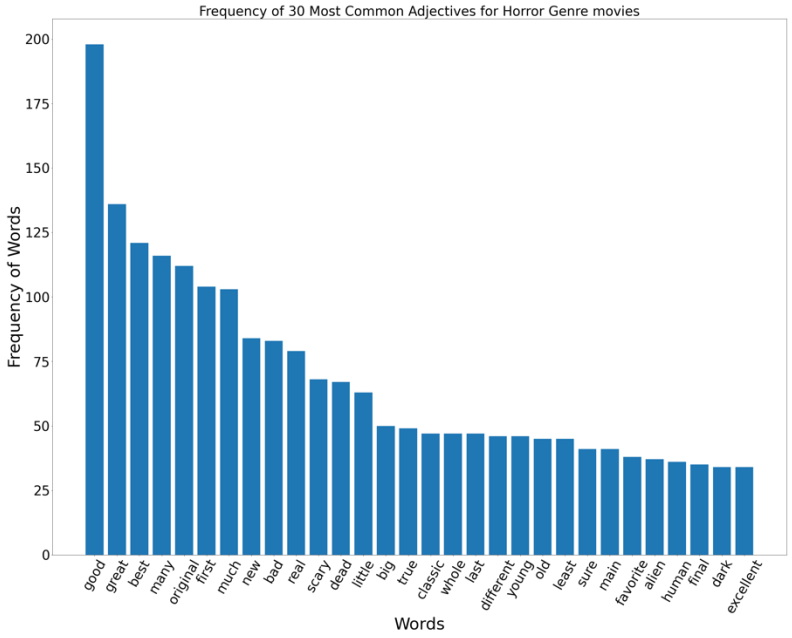

Fig. 5. Top 30 most frequent words shown for Horror Genre

3.3.6 Aspect-based on genre

The following graphs are plotted using a lexicon base package called Spacy. Spacy classifies lemmatized text into nine different categories and the most frequently occurring words are collected for each genre. Representative actors and directors for each genre appear, such as a famous romance actress "Audrey Hepburn" for Romance and a famous horror movie director "James Carpenter" for Horror. It also shows representative places such as "Japan," "New York," and "France" for Romance and "Texas" for Horror. This is doing much better than previous EDA conducted on general EDA in terms of bringing out the separation among different aspects, which leads to better sentiment analysis.

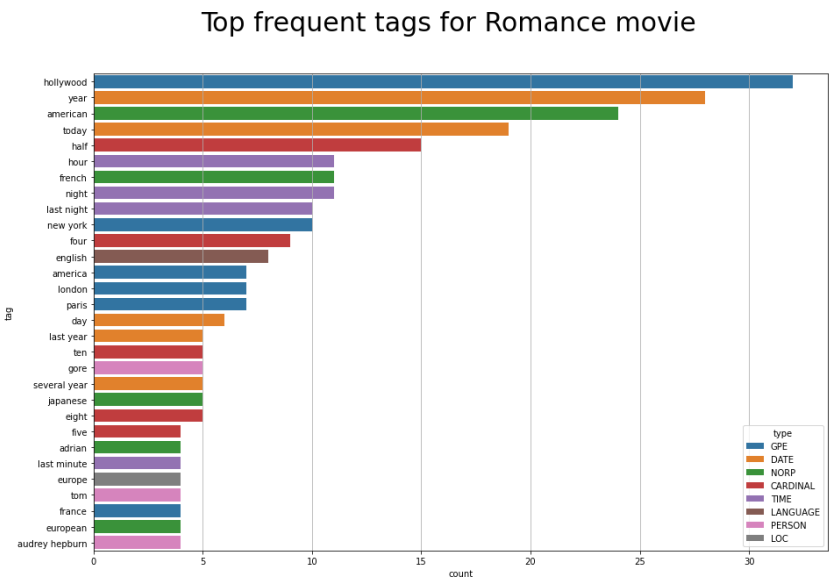

Fig. 6. Top 30 most frequent tags shown for Romance Genre using Spacy

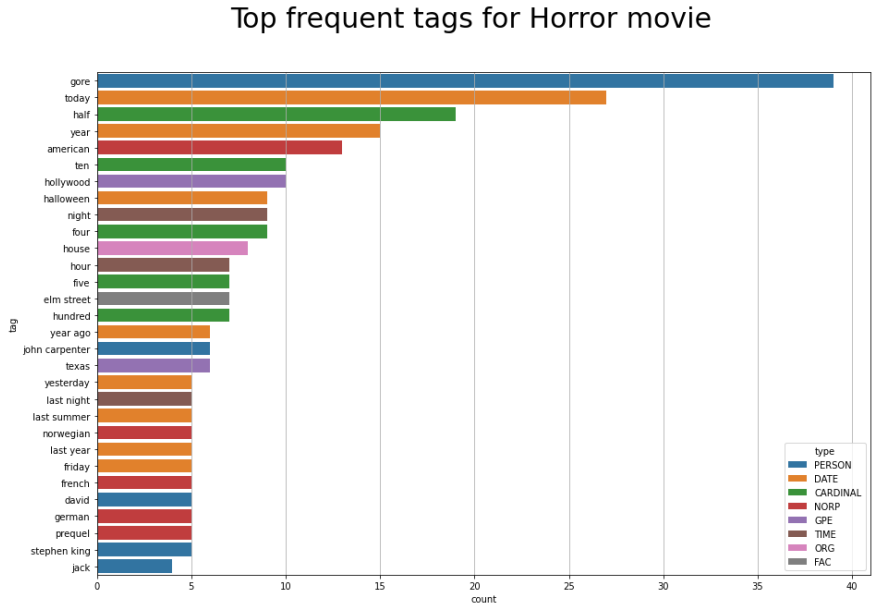

Fig. 7. Top 30 most frequent tags shown for Horror Genre using Spacy

3.3.7 Latent Dirichlet Allocation (LDA) analysis

LDA is an unsupervised technique that can automatically identify topics among given text. Gensim and pyLDAvis are the packages used to construct this visualization. Clusters are determined through trial and error and three clusters visually appear to be the optimal number of topics as in **Fig. 8** below. The bar chart on the right side shows the term frequency for both within the document and within the selected group. Further analysis can be done using the sensitivity parameter lambda in the visualization tool.

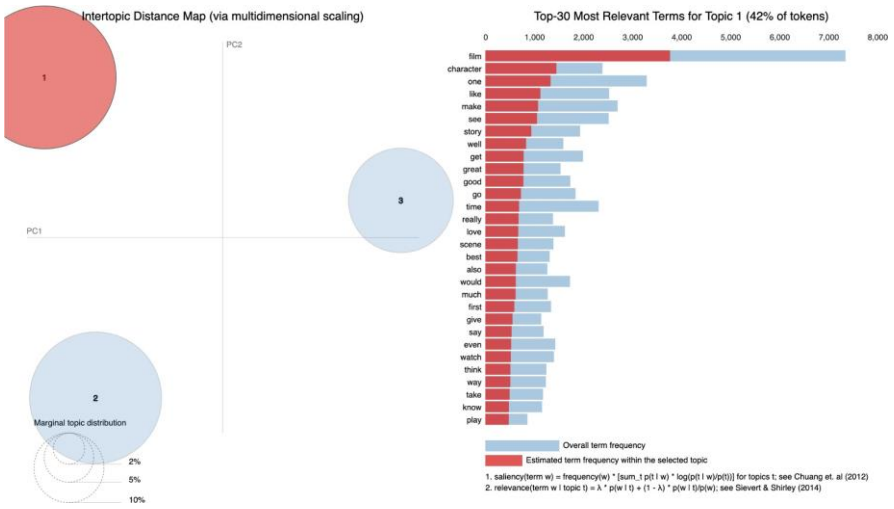

Fig. 8. Interactive plot for LDA result shows group of 3 topics is the most optimal.

4 Methods

4.1 Aspect extraction

Aspect extraction is a technique used to extract information from unstructured text and cluster them into predefined categories such as person, location and organization [28]. It is extremely useful when it comes to granular sentiment analysis of exploring different aspects within user generated content. The following two feature extraction methods in Figure 9 will be discussed, employed for aspect extraction. The extracted aspect words are aggregated onto a lexicon list that will be later used to split the text into various aspects.

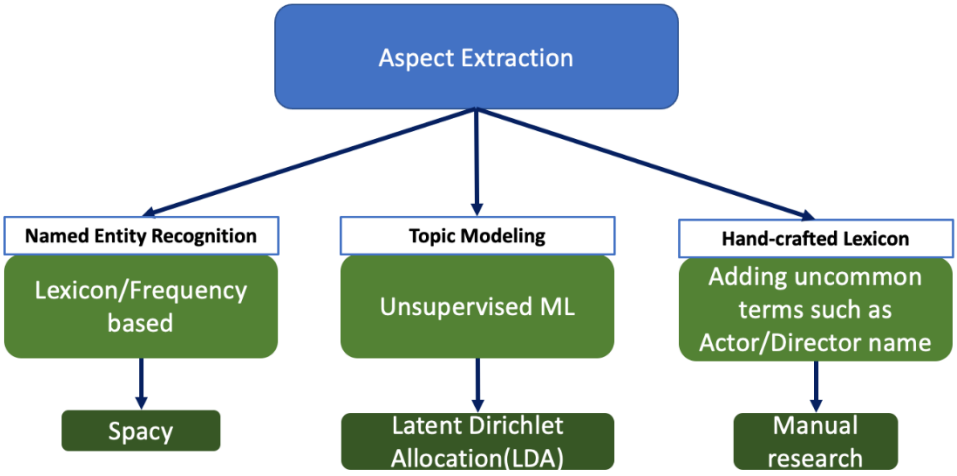

Fig. 9. Aspect extraction implemented in three different methods.

4.1.1 Rule-based (Spacy)

Spacy is one of the best open source NER tools available and provides several predefined entity categories. After each word is tagged, only the nouns that are semantically similar to the predefined aspects using word2vec will be extractedand stored in a bag of words for each aspect.

4.1.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a widely used unsupervised topic modeling technique that trains the data with multinomial distribution with user defined K latenttopics [33]. To handle overfitting, LDA uses Dirichlet prior to the parameters of the topic and unigram distributions. Different K latent topic numbers will be run in conjunction with different machine learning tools and the one with the highest accuracy will be selected.

4.1.3 Hand-crafted Lexicon

Uncommon terms such as actor or director's names were manually searched online and added to the lexicon. This process ensures that the directing and acting aspect based components filter the terms that are associated them.

4.2 Text Vectorization

Each feature extraction method listed above will generate aspects that are different in content and scale. By aggregating the resulting aspects, each review is reduced to the dominant aspect specific chunk. Then the textual data has to be converted into vectors in order to fed into machine learning algorithms. Term Frequency Inverse Frequency(TF-IDF) and CountVectorizer perform the essential task of text vectorization in which it associates each word with a number that represents how relevant that word is in the document. TF-IDF differs from the CountVectorizer in that TF-IDF considers the overall document weightage of a word and penalizes the most frequent words. Both methods will be implemented and compared for its performance against the model accuracy.

4.3 Model Building

After the reviews are split and reduced by the dominant aspect specific for each review, text information is converted into vectors using both TF-IDF and CountVectorizer described above. Different machine learning tools will be applied and conducted normal training/testing on those vectors. Then they are classified into sentiment of -1 and 1using machine learning tools. Genre driving factors are taken into consideration when determining the final sentiment prediction. All the steps leading up to sentiment analysis using machine and deep learning are displayed in Figure 10 below.

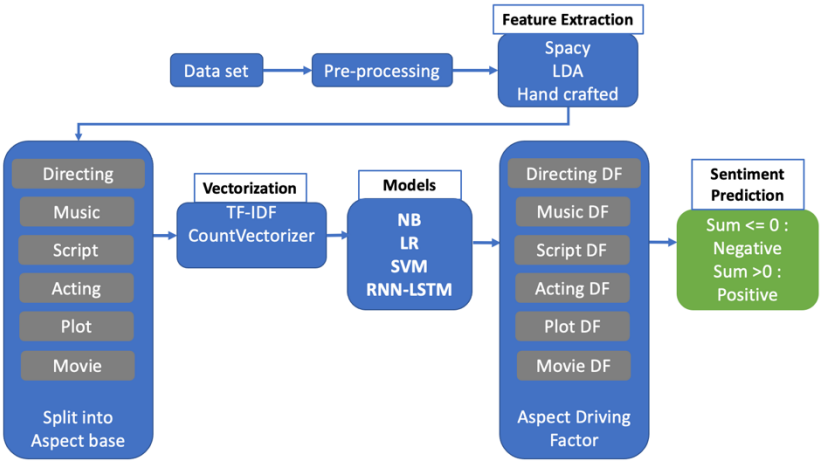

Fig. 10. General workflow leading up to sentiment using various feature extraction and machine and deep learning tools. LDA=Latent Dirichlet Allocation, LR=Logistic Regression, NB=Naive Bayes, SVM=Support Vector Machine, RNN-LSTM=Recurrent Neural Network-Long Short-Term Memory

Several machine and deep learning models such as Naïve Bayes, Logistic Regression, Support Vector Machine and LSTM of Recurrent Neural Network will be used to train and test the review content and predetermined sentiments.

5 Results

5.1 Related results from prior publications

The use of weighted driving factors to identify movie aspects in previous work was particular in that they gave highest importance to certain aspects such as "movie," “acting,” and “plot” rather than equal weights as this resulted in a slight drop of metric performance. Using weighted aspects also worked well to suppress opinions in the reviews on other factors, such as when reviews were longer in length which added to its sentiment influence. Accuracies also varied between genres ranging from 63.3% to 87.3% which was for the “crime” genre. Driving factor importance also shifted between genres, “crime” for example found “movie”, “screenplay” and “plot” to be more influential. The genre specific accuracies were found to increase when reviewers had made comments on those driving factors [31]

Extraction of aspects was found to be most successful using a frequency-based TD-IDF approach in prior research, though this was limited by the necessity for very large amounts of data to achieve this. A rule-based approach worked best in prior publications for extraction at 92% precision [42]. However, this also typically required a large set of defined hand crafted rules which would tend to perform badly in any undefined instances and specifically in cases of named entity recognition for different languages [32]

5.2 Model Results

The hypothesis of this research is that assigning high driving factors to certain aspect of the movie result in higher accuracy of the models. The driving factors are randomly assigned to various movie aspects and their impact tied to each aspect and genre leading to the sentiment classification has been fully investigated based on the

accuracy of each model. Figure 11 shows the distribution of average aspect driving factors for each genre for the highest accuracy observed in this research. After numerous iterations to ensure the generalization of the driving factors over the entire dataset using cross validation, the study revealed that assigning higher driving factors to the plot and acting for Action, directing and music for Horror, screenplay and acting for Comedy, screenplay and acting for Romance and plot and directing for Sci-fi genre movies, result in the higher accuracy of the models that were utilized in this research. Figure 12 shows the generalized result of top contributing movie aspects per genre group based on the model performance.

```
The highest accuracy observed is: 67.1971706454465 %
----------------------------------------------------------------------
The average driving factors for the 'Action' movies are:
      DC    |     SC     |     MC    |     PC     |     AC    |     MvC
[0.50242191 0.47008948 0.49001913 0.48296697 0.5143432  0.48698634]
----------------------------------------------------------------------
The average driving factors for the 'Horror' movies are:
      DC    |     SC     |     MC    |     PC     |     AC    |     MvC
[0.46500126 0.52645268 0.48873629 0.50518756 0.44750295 0.48104021]
----------------------------------------------------------------------
The average driving factors for the 'Comedy' movies are:
      DC    |     SC     |     MC    |     PC     |     AC    |     MvC
[0.4832768  0.49142149 0.51031646 0.5280442  0.49388201 0.47653467]
----------------------------------------------------------------------
The average driving factors for the 'Romance' movies are:
      DC    |     SC     |     MC    |     PC     |     AC    |     MvC
[0.52161601 0.51218858 0.49780425 0.48048248 0.50146239 0.46147347]
----------------------------------------------------------------------
The average driving factors for the 'Sci-Fi' movies are:
      DC    |     SC     |     MC    |     PC     |     AC    |     MvC
[0.48974328 0.54835633 0.46172844 0.54758009 0.48752774 0.52973507]
----------------------------------------------------------------------
```

Fig. 11. Driving factor distribution among various aspects and genre

	1st	2nd	3rd
Action	PC	AC	DC
Horror	DC	MC	SC
Comedy	SC	AC	DC
Romance	SC	AC	MC
Sci-fi	PC	DC	AC

Fig. 12. Driving factor distribution among various aspects and genre

10-Fold Cross Validation (CV) was performed on the entire dataset to measure the model effectiveness. In Figure 13 it shows model performance when TF-IDF vectorization was used. SVM performed the best when using TF-IDF vectorization in terms of the average accuracy and standard deviation across its CV models. Naïve Bayes performed the worst having to ingest TF-IDF vectors, which are numbers that are continuous rather than discrete. When using CountVectorizer, the performance of LR and SVM dropped by 3-4%. In Figure 14, it shows that NB performs the best in terms of average accuracy and standard deviation among its CV result. This shows that pairing vectorizer with models properly is imperative in performance. The research proceeded with using CountVectorizer for NB and TF-IDF for LR and SVM models.

	NB	LR	SVM
1	67.64	66.14	66.14
2	65.517	64.28	65.34
3	65.87	64.46	66.4
4	65.61	67.81	66.49
5	65.25	66.4	66.75
6	64.63	65.25	67.11
7	66.4	65.43	64.98
8	66.14	64.63	66.22
9	64.19	66.41	66.61
10	67.34	66.05	67.19
Average	65.8587	65.686	66.323
SD	1.0843806	1.0921253	0.7058179

Fig. 13. 10-Fold CV accuracy compared among ML models used with TF-IDF vectorizer

	NB	LR	SVM
1	67.64	64.63	65.43
2	65.517	63.39	64.01
3	65.87	64.37	64.37
4	65.61	61.54	62.07
5	65.25	64.28	64.28
6	64.63	64.72	66.67
7	66.4	63.13	64.01
8	66.14	64.19	65.34
9	64.19	63.66	64.36
10	67.34	64.37	64.99
Average	65.8587	63.828	64.553
SD	1.0843806	0.9605531	1.2000653

Fig. 14. 10-Fold CV accuracy compared among ML models used with CountVectorizer

Figure 15 shows the result of running the models without accounting driving factors. The result proves that incorporating driving aspect and genre driving factors increase the accuracy on average 3 to 4%. Incorporating driving factors resulted in the highest observed sentiment prediction accuracy of 68%, compared to 63% without using them. Thus, the hypothesis of this research is reasonably accepted and can be further developed by refining lexicon base and delving into deep learning models.

	NB	LR	SVM
1	63.66	63.66	65.78
2	62.7	61.62	63.04
3	64.45	63.04	63.83
4	63.04	64.1	65.16
5	65.78	64.98	63.39
6	63.12	62.59	62.95
7	63.12	65.16	63.39
8	64.1	62.68	62.59
9	61.8	62.24	61.45
10	64.36	63.04	62.42
Average	63.613	63.311	63.4
SD	1.1119957	1.1556667	1.2779411

Fig. 15. 10-Fold CV accuracy result with no driving factors accounted. NB with CountVectorizer, LR and SVM with TF-IDF vectorizer

6 Discussion

Evaluating the user generated reviews on a large scale is crucial in understanding the business performance and unlocking invaluable insights that can move the needle forward. Aspect based sentiment analysis is one of the most widely discussed topics in the Natural Language Processing (NLP) community. The insights on which aspect and genre driving factors drive the accuracy of the sentiment prediction can greatly assist with the movie industry and its stakeholders.

After reviewing the previous research papers on our topic of research, it was found that most researchers focused on the general breakdown of the aspect-based sentiment analysis for movie reviews while only a few emphasized on the predictive capability of ABSA models. To contribute to the existing research and to improve model performance, existing problems surrounding ABSA in movie reviews have been investigated and examined to build several recommendation models. Aspect and genre driving factor assisted models presented in this research allowed understanding in how granular level sentiment analysis can be performed. This can be beneficial to recommendations in movies but also to other businesses that utilize customer reviews to enhance certain aspects of their business.

For example, restaurant chains can actively train and utilize these aspect-based models with their online reviews to fine tune their key performance indicators (KPIs) by compartmentalizing their overall ratings into service, food, price, and atmosphere. The results can extract actionable items and deliver direct impacts to the business and its stakeholders.

6.1 Ethics

In section 3, the dataset was described with its extraction method. The ethical scope of this project was to scrape the data from IMDB website properly. It is acknowledged that data was scrapped without permission from IMDB, however, the authors of this research paper do not seek any monetary benefit or intend to commercialize applications in the future, thus, there are no ethical issues of concern.

There are a few other biases that need to be addressed in this dataset, such as population distribution and nationality. Even though the dataset seems to be limited to the movies that were released in United States, the opinions gathered are international. It is also not clear when and how long the audience was exposed to the movies, and this might have affected the movie reviews due to generational differences. Gender was also largely more representative of male reviewers as was seen to be common throughout this type of media. To decrease this form of bias the site with the highest female representation was selected for data collection. The dataset was randomly sampled and is therefore treated as non-biased for the purpose of this study.

Furthermore, statistical methods were deployed in transforming the dataset to a more suitable form for a proper EDA analysis to ensure the dataset is clean. To achieve the aim and objectives of the research, machine learning models were used to perform the following tasks.

- Exploratory analysis on the dataset to make statistical inferences, this method also shows the distribution of the reviewers' sentiments with their focused aspects.
- Statistical models were built and trained using based on the cleaned-up dataset and the models with the highest accuracy were selected while rules were set aside to make sure there is no overfitting or underfitting to build a better recommendation system.

- It is discovered that some of these aspects are more important than others. These aspects are important as they can help to easily know reviewers' areas of concentration.

6.2 Challenges and Limitations

There were several challenges that needed to be addressed over the course of the research. As noted in section 2.6 and 6.1, securing a dataset that is unbiased and representative of the general audience was a concern. The dataset generated from the general audience rather than professional movie critics was targeted due to the public audience being the main revenue stream and the overall sentiment of the movie. During the process of identifying the right dataset for this study, the lack of information on demographics, scoring metric and nationality of the reviewers were hindering factors in fully grasping the data. However, a few general assumptions were established to mitigate the bias as noted in section 2.6.

Another challenge that was encountered during the study was splitting the text into different aspects properly with the established lexicon base. Part Of Speech (POS) tagging was utilized to filter out adjectives, verbs, nouns, and their combinations that belong to the lexicon base. Even though a supervised learning and a traditional lexicon method (SPACY) was utilized to establish a lexicon base for the input text, it appeared that not every contributing word to the sentiment was listed on the lexicon base. This could be a future work to see how models perform with a refined lexicon base that inclusively covers the input text words.

6.3 Future work

For future work, it is recommended to extend this work to include a time dependent lexicon base. Time plays an important role in changing sentiments on how things are perceived by the public. The article about how the public sentiment on the conservation planning in a local community evolves over time addresses the need to consider time as an important factor in sentiment analysis [11]. The scale of News coverage changes from local tensions to raising awareness of the importance of iconic species to the global domain. These events are strongly affected by natural and social events [11]. Likewise, as a movie grows in popularity from the local community to the world over time and as it receives more acceptance/criticism all aroundthe world, the review content will dynamically change in conjunction with social and historical events. There are historical semantic shifts and changes happening to many words [21]. Detecting meaning changes can help clarify the ambiguity of dynamical linguistic systems.

As mentioned in section 6.2, establishing a solid lexicon base for the purpose of filtering out aspect words into various aspects of the movie is imperative. Striking a balance between under-filtering and over-filtering aspect words would significantly increase accuracy of the models in the future.

The supervised learning models used in this study do not consider the inter word meaning dependencies and the nuanced context in which the words are used. Even though aspect and genre driving factors are already incorporated, it will be beneficial to examine the effect of including inter word dependencies in the model as a feature.

A final arena that this study can further developed upon would be deep learning applications. This study also includes the Recurrent Neural Network (RNN) Long Short Term Memory (LSTM) to apply deep learning towards the sentiment analysis. As seen Figure 16, the model did not perform well compared to supervised learnings conducted in this study, however, the root cause of the discrepancy with other supervised learnings can be fully examined in the future. Additionally, the deep learning application can be improved upon tuning parameters and refining neural

network environment in the future.

	RNN LSTM
1	51.28
2	51.81
3	49.95
4	52.25
5	51.72
6	53.41
7	50.23
8	51.72
9	51.71
10	52.69
Average	51.677
SD	1.0335597

Fig. 16. 10-Fold CV accuracy result with LSTM

7 Conclusion

The project was conducted to find which movie aspects drive the sentiment of the reviews using different driving factors. Four different sentiment classification methods were utilized to find the sentiment: three supervised machine learning of Logistic Regression, Naive Bayes, Support Vector Machines, and one deep learning of Recurrent Neural Network. Despite challenges of establishing proper lexicon list for the aspect related words, the added feature of aspect and genre driving factor boosted accuracy of the aspect-based models, thus reasonably improved the predictions of the review sentiment. The study results also suggest that these driving factor assisted models can deliver insights on which aspects under certain genre drive the most sentiment for any unseen test dataset.

The models presented here provide a framework for various other applications. As discussed earlier, the study results presented here are not only useful in the movie industry but also useful in other industries in which their business is driven by user generated reviews. Analysis of reviews incorporating driving factors can increase model accuracy while providing insight into customer motivations and concerns.

Acknowledgments.

Faizan Javed, PhD. - SMU Professor and Advisor
Jacquelyn Cheun, PhD. – SMU Capstone Professor

References

1. Akter, Shariar; McGarthy, Grace; Sajib, Shariar; Michael, Katina; Dwivedi, Yogesh; Ambra, John; Shen, K. "Algorithmic bias in data-driven innovation in the age of AI" International Journal of Information Management Volume 60, October 2021, 102387. https://www-sciencedirect-com.proxy.libraries.smu.edu/science/article/pii/S0268401221000803
2. Ali, Nehal & Hamid, Marwa & Youssif, Aliaa. (2019). "Sentiment Analysis For Movies Reviews Dataset Using Deep Learning Models". International Journal of Data Mining & Knowledge Management Process. 09. 19-27. 10.5121/ijdkp.2019.9302. https://www.researchgate.net/publication/333607586_SENTIMENT_ANALYSIS_FOR_MOVIES_REVIEWS_DATASET_USING_DEEP_LEARNING_MODELS
3. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). "Learning Word Vectors for Sentiment Analysis." *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).* http://ai.stanford.edu/~amaas/data/sentiment/
4. Atif Khan, Muhammad Adnan Gul, M. Irfan Uddin, Syed Atif Ali Shah, Shafiq Ahmad, Muhammad Dzulqarnain Al Firdausi, Mazen Zaindin, (2020) "Summarizing Online Movie Reviews: A Machine Learning Approach to Big Data Analytics", Scientific Programming, vol. 2020, Article ID 5812715. https://doi.org/10.1155/2020/5812715
5. Baid, Palak & Gupta, Apoorva & Chaplot, Neelam. (2017). "Sentiment Analysis of Movie Reviews Using Machine Learning Techniques". International Journal of ComputerApplications. 179. 45-49. 10.5120/ijca2017916005. https://www.researchgate.net/publication/321843804_Sentiment_Analysis_of_Movie_Reviews_using_Machine_Learning_Techniques'
6. Battaglia, James, "Everyone's a Critic: Film Criticism Through History and Into the Digital Age" (2010). Senior Honors Theses. 32. https://digitalcommons.brockport.edu/honors/32
7. Brar, Gurshobit; "Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 16 (2018) pp. 12788-12791. https://www.ripublication.com/ijaer18/ijaerv13n16_53.pdf
8. Cambria, E; Schuller, B; Xia, Y; Havasi, C (2013). "New avenues in opinion mining and sentiment analysis". IEEE Intelligent Systems. 28 (2): 15–21. CiteSeerX 10.1.1.688.1384. https://doi.org/10.1109%2FMIS.2013.30
9. Coggan, D. (2016, June 23). Male film critics greatly outnumber female critics, study finds. EW.Com. https://ew.com/article/2016/06/23/film-criticism-gender-study/
10. Collazo, M. (2014, April 30). How Movie Critics and Moviegoers View Films Differently. The Artifice. https://the-artifice.com/movie-critics-and-moviegoers-view-films-differently/
11. Ernoul,Lisa; Wardell,Angela (2016) "Representing the Greater Flamingo in Southern France: A semantic analysis of newspaper articles showing change over time". *Ocean and Coastal Management* Vol 133 pg 105-113 https://www.sciencedirect.com/science/article/abs/pii/S0964569116302101
12. Fang, X., Zhan, J. (2015) "Sentiment analysis using product review data", *Journal of Big Data 2,5*. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2
13. Gruhl, D., R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2005. "The predictive power of online chatter." KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 78-87, August. doi: 10.1145/1081870.1081883. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.8553&rep=rep1&type=pdf
14. Haddi, Emma; Liu, Xiaohui; Shi, Yong. (2013) "The Role of Text Pre-processing in Sentiment Analysis", Procedia Computer Science: Vol 17, Page 26-32 https://www.sciencedirect.com/science/article/pii/S1877050913001385
15. Hatzivassiloglou, V; McKeown, K, 1997. "Predicting the Semantic Orientation of Adjectives." 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, July, pp. 174-181. https://www.aclweb.org/anthology/P97-1023/
16. IMDb | Help. (n.d.). IMDB. Retrieved August 29, 2021, from https://help.imdb.com/article/imdb/track-movies-tv/ratings-

faq/G67Y87TFYYP6TWAV?ref_=helpms_helpart_inline#
17. IMDb vs Rotten Tomatoes: The Wisdom of Crowd Goes to The Movies. (2018, November 28). Wordpress. https://learncuriously.wordpress.com/2018/11/25/wisdom-of-crowd-goes-to-the-movies/
18. Kang, Hanhoon ; Yoo, Seong Joon ; Han, Dongil (2012) "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", *Elsevier Ltd Expert systems with applications*, Vol.39 (5), p.6000-6010 https://www-sciencedirect-com.proxy.libraries.smu.edu/science/article/pii/S0957417411016538
19. Khaleghi, Ryan; Cannon, Kevin; and Srinivas, Raghuram (2018) "A Comparative Evaluation of Recommender Systems for Hotel Reviews," *SMU Data Science Review*: Vol. 1 : No. 4 , Article 1. https://scholar.smu.edu/datasciencereview/vol1/iss4/1
20. Kilkenny, K. "How the Internet Led to the Decline of Female Film Critics". The Atlantic. 2015-12-27. Retrieved 2018-06-21 https://www.theatlantic.com/entertainment/archive/2015/12/female-film-critics/421629/
21. Kulkarni, Vivek; Perozzi, Bryan; Al-Rfou, Rami; Skiena,Steven "Statistically Significant Detection of Linguistic Change"(2020) *Github* http://vivekkk.github.io/langchangetrack/data/kulkarni.pdf
22. Lakshmi Devi B., Varaswathi Bai V., Ramasubbareddy S., Govinda K. (2020) Sentiment Analysis on Movie Reviews. In: Venkata Krishna P., Obaidat M. (eds) Emerging Research in Data Engineering Systems and Computer Communications. Advances in Intelligent Systems and Computing, vol 1054. https://doi.org/10.1007/978-981-15-0135-7_31
23. Lee, Sung-Won ; Jiang, Guangbo ; Kong, Hai-Yan ; Liu, Chang(2020), " A difference of multimedia consumer's rating and review through sentiment analysis", *Multimedia tools and applications* https://link-springer-com.proxy.libraries.smu.edu/article/10.1007/s11042-020-08820-x
24. Ligthart, A., Catal, C. & Tekinerdogan, B. (2021) "Systematic reviews in sentiment analysis: a tertiary study," *Artif Intell Rev*. https://doi.org/10.1007/s10462-021-09973-3
25. Lochmiller, Chase; "A Survey of Techniques for Sentiment Analysis in Movie Reviews and Deep Stochastic Recurrent Nets", Department of Computer Science Stanford University, Reports 2016. https://cs224d.stanford.edu/reports/chase.pdf
26. Mamtesh, Seema Mehla (National Institute of Technology Kurukshetra, India) "Sentiment Analysis of Movie Reviews using Machine Learning Classifiers", *International Journal of Computer Applications (0975–8887) Volume 182 – No. 50, April 2019.* https://www.ijcaonline.org/archives/volume182/number50/mamtesh-2019-ijca-918756.pdf
27. Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed (2018) "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," *SMU Data Science Review*: Vol. 1 : No. 4 , Article 7. https://scholar.smu.edu/datasciencereview/vol1/iss4/7
28. Oswal,Sangeeta; Soni,Ravikumar, Narvekar, Omka (2019) "Named Entity Recognition and Aspect based Sentiment Analysis", *International Journal of Computer Applications*: Vol 178 - No 46 https://www.ijcaonline.org/archives/volume178/number46/30859-2019919367
29. Ortony, Andrew; Clore, G; Collins, A (1988). The Cognitive Structure of Emotions (PDF). Cambridge Univ. Press. http://www.cogsci.northwestern.edu/courses/cg207/readings/Cognitive_Structure_of_Emotions_excerpt.pdf
30. Pang, B., Lee, L. 2004. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts." Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pp. 271-278, July. https://www.aclweb.org/anthology/P04-1035/
31. Parkje, Viraj; Biswas, Bhaskar; "Aspect Based Sentiment Analysis of Movie Reviews" (2014) *International Conference on Soft Computing & Machine Intelligence* https://ieeexplore.ieee.org/document/7079348
32. Patil,Nita; Patil, Ajay; Pawar, B.V. (2019) "Named Entity Recognition using Conditional Random fields", *International Conference on Computational Intelligence and Data*

Science(ICCIDS): Vol 167 p. 1181-1188
https://www.sciencedirect.com/science/article/pii/S1877050920308978

33. Polifroni,Joe; Mairesse,Francois (2011) "Using Latent Topic Features for Named Entity Extraction in Search Queries", *Interspeech*
http://farm2.user.srcf.net/research/papers/is11-lda.pdf
34. Ramya, V.Uma; "Sentiment Analysis of Movie Review using Machine Learning Techniques" International Journal of Engineering & Technology, 7 (2.7) (2018) 676-681.
https://www.sciencepubco.com/index.php/ijet/article/view/10921
35. Reza Maulana et al; (2020), "Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain" J. Phys.: Conf. Ser. 1641 012060.
https://iopscience.iop.org/article/10.1088/1742-6596/1641/1/012060
36. Reynolds, M. (2017, October 24). You should ignore film ratings on IMDb and Rotten Tomatoes. WIRED UK.
https://www.wired.co.uk/article/which-film-ranking-site-should-i-trust-rotten-tomatoes-imdb-metacritic
37. Sankar, H., Subramaniyaswamy, V. 2017. "Investigating sentiment analysis using machine learning approach." International Conference on Intelligent Sustainable Systems (ICISS), IEEE, pp. 87-92, December 7-8.
https://ieeexplore.ieee.org/abstract/document/8389293/
38. Somya Dwivedi, Harsh Patel and Shweta Sharma; "Movie Reviews Classification Using Sentiment Analysis", Indian Journal of Science and Technology, Vol 12(41), November 2019.
https://dx.doi.org/10.17485/ijst/2019/v12i41/145554
39. Swathi H., S. S. Aravinth, V. Nivethitha, T. Saranya, R. Nivethanandhini; "Sentiment Analysis of Movie Review using data Analytics Techniques", MAR 2019, IRE Journals, Volume 2 Issue 9.
https://irejournals.com/formatedpaper/1701029.pdf
40. Wollmer, M ; Weninger, F ; Knaup, T ; Schuller, B ; Congkai Sun ; Sagae, K ; Morency, L-P(2013) "Youtube Movie Reviews: Sentiment Analysis in an Audio-Visual Context", *IEEE intelligent systems*, Vol.28 (3), p.46-53
https://ieeexplore-ieeeorg.proxy.libraries.smu.edu/stamp/stamp.jsp?tp=&arnumber=6487473
41. Yao, Yao; Angelov, Ivelin; Rasmus-Vorrath, Jack; Lee, Mooyoung; and Engels, Daniel W. (2018) "Yelp's Review Filtering Algorithm," *SMU Data Science Review*: Vol. 1 : No. 3 , Article 3.
https://scholar.smu.edu/datasciencereview/vol1/iss3/3
42. Yadav, Kaustubh; (2006) "A Comprehensive Survey on Aspect Based Sentiment Analysis", School of Computer Science and Engineering- SCOPE, Vellore Institute Of Technology.
https://arxiv.org/abs/2006.04611